RAND

Alignment Among Secondary and Post-Secondary Assessments in Oregon

Vi-Nhuan Le, Abby Robyn

DRU-2528-EDU

March 2001

RAND Education

By Vi-Nhuan Le and Abby Robyn

Alignment Among Secondary and Post-Secondary Assessments in Oregon

Background: The Current Role of Assessments in Secondary and Post-Secondary Schooling

A Multitude of Assessments

As students progress through high school and into institutions of higher education, they take numerous tests that vary in scope, content, and purpose. At the K-12 level, almost all of the states are currently using or developing assessments that are aligned with state standards (American Federation of Teachers, 1999). In addition to the K-12 tests, students who plan to attend college usually take another bevy of assessments, ranging from nationally administered college admissions exams to university-specific course placement tests.

Assessments, as "snapshots" of student performance, are imperfect, indirect measures of what students have accomplished. Nevertheless, they can provide valuable information about a student's capabilities within particular contexts. A well-constructed English composition placement test, for instance, can help a university determine whether a student is ready for college-level writing or needs a remedial composition course. A statewide achievement test, on the hand, typically has no direct relevance to college placement or admission. Rather, the aims of most state tests are to sample from a range of topics and to cover material learned across several grades so that state educators, policymakers, parents, and citizens can form a better understanding of how their students and schools are performing, and whether or not they are improving. Consequently, K-12 and college tests may not resemble one another closely in the constructs that they measure. In other words, they may not be well aligned.

Discrepancies among assessments, however, should not necessarily be construed as problematic. Especially when tests reflect dissimilar curricular frameworks, or are designed for different uses, we would expect the measures to vary. A college admissions test that is intended for prospective college-bound examinees may need to include

different kinds of items than a K-12 test that is taken by students whose proficiency levels are more disparate. Thus, efforts to make measures more consistent are not always desirable. This point will be more fully explored in later sections.

This Study's Goal: Determining the Extent to which Assessments are Aligned

The goal of the present study is to investigate the degree of alignment among different types of tests in five case study states, and to explore the consequences of any discrepancies. We compare assessments used for college admissions, placement, and K-12 system monitoring and accountability in each state, classifying items along several dimensions. As a result, we obtain for each state a summary of the ways in which the assessments are and are not aligned with one another, and discuss possible implications. In particular, we discuss the signaling functions of the tests, which is the primary motivation for this alignment study. That is, we examine the potentially confusing messages that students may receive from the discrepant assessments.

Although our analysis will document many examples of misalignments among the assessments, we do not necessarily advocate uniformity across all the measures. Indeed, there are many instances in which discrepancies among the exams are necessary. Instead, our principal purpose is to examine the signals that the tests send.

This analysis is part of a larger study examining the transition from secondary to post-secondary education (Kirst, 1998). The importance of this transition is underscored by a 1995 National Education Association survey in which 82% of House and Senate Education chairs polled viewed the improvement of connections between colleges and schools as among their highest priorities for higher education (Edgerton, 1997).

Structure of this Oregon Case Study

This case study is one of five that have been developed for the following states: California, Georgia, Maryland, Oregon, and Oregon. Before we present the Oregon case study, we discuss the study's methodology and limitations, and then describe the problems and benefits related to the discrepancies between different assessments. We also discuss the college admissions tests that are common to all five states, and provide a brief description of the other kinds of assessments we analyzed. This is followed by the

Oregon case study. We provide for this study, as we do for the others, a description of the testing environment, the assessments specific to the particular state, and the results of the alignment analyses.

Study Methodology

The alignment analysis involved two major phases. In phase 1, we developed a framework of specifications for each subject. We examined several existing assessment frameworks, such as those used to develop the National Assessment of Educational Progress (NAEP), and combined them to produce a set of specifications that addressed the range of topics and item types appearing on the tests included in this study. We then applied these frameworks to our set of tests, and made several rounds of modifications in response to difficulties we encountered in conducting the alignment exercise. The process was similar to one that is often used for developing scoring rubrics for open-ended assessment items. The resulting frameworks are described later in this report.

Phase 2 consisted of the actual alignment exercise. Two raters who had expertise in both the relevant subject area and in the application of scoring criteria to assessment results conducted the alignment for each subject. Initially, both raters independently coded several of the math and English assessments, and compared their results. When raters differed in their interpretations of the framework components, they discussed the difference until agreement was reached. The raters continued coding both math and English exams until they were satisfied that the scoring guidelines in both subject matters were being applied in a consistent manner. Thereafter, one rater coded all the remaining English assessments, and the other rater coded the remaining math tests.

For the assessments that both raters coded, there was a high level of agreement (percentages of approximately 85%-100%) for most categories. Two exceptions were content area in math, where items often assessed skills in more than one area, and passage topic in reading, because passages often could be coded as addressing more than one topic. A final exception was the cognitive process category in math, discussed further below. For these categories, agreement tended to be approximately 70%.

The use of expert judgments is a fairly common approach to studying alignment as well as content validity (e.g., Sireci, 1998). The evidence gathered through this study

will be useful in evaluating the validity of currently used tests in terms of the purposes for which they were designed. However, this study does not provide a complete picture of these assessments, and other analytic approaches might lead to somewhat different conclusions. An analysis of scores on a state's tests might reveal, for example, that seemingly different instruments rank order students roughly equivalently. Similarly, observations and interviews with students as they take the tests, an approach that is sometimes used during the test development process, could result in somewhat different interpretations of a test's reasoning requirements. Finally, increasing the number of forms studied for each assessment would enhance the generality of our findings. The studied tests represent a sample of skills from a single testing occasion, and forms from other occasions will vary somewhat. This is especially true when we analyze alignment among English Language Arts (ELA) topics, where any given test form provides a limited sample (e.g., there may be only one essay).

The Importance of Aligning Assessments

For many promoters of test-based accountability systems (Achieve, Inc., 1999) and advocates of systemic reform (e.g., Smith & O'Day,1991), aligning various aspects of the educational system to support a set of common goals is an essential part of educational reform efforts. Especially important to standards-based reform efforts is the degree to which the standards and the assessments used to measure progress toward the standards are consistent with one another. Standards and assessments that are not aligned with one another or that encourage a focus on low-level skills may create mixed messages and confusion for students, teachers, and others involved in promoting student learning. Here we list three major concerns with discrepancies among tests.

(1) First, the content of test items sends messages to the students who take them. Discrepancies are likely to occur when two tests reflect different philosophies concerning what students should know and be able to do. For instance, many of the discrepancies among K-12 and university-level tests result from reforms that have taken hold at one level of the educational system but not another. This is particularly

true in states where new tests have been developed to reflect state standards or frameworks. In such cases, the skills and knowledge advanced by the statewide exam may be at odds with those emphasized by the college placement tests. This creates an unclear set of signals for students and teachers regarding the kinds of skills that are valued. If students enter college unaware of what skills they will be expected to demonstrate, their performance may suffer.

(2) **A second important aspect of alignment (and discrepancy) involves the consistency (and inconsistency) with which students are rank ordered or classified into categories or programs (e.g., remedial instruction) by different tests.** If two tests are designed to measure the same abilities, evidence must be gathered to show that students who do well on one test tend to do well on the other. Although most tests of academic achievement tend to correlate highly with one another, even when subject and item format differ, it is nonetheless important to evaluate the magnitude of this correlation and the consistency of any classification that results from test use.

(3) **Third, it is essential that the cut scores used for decision-making be comparable across assessments and set in a technically sound and credible manner.** The placement process often involves selecting a cut score on an exam and assigning students to programs or courses based on whether or not their scores exceeded this cut score. Statewide assessment programs are increasingly reporting student performance in terms of benchmarks similar to the achievement levels used on the National Assessment of Educational Progress (NAEP). These efforts have been criticized (see, e.g., Burstein et al., 1995/96; National Academy of Education, 1993), in part because the process of mapping performance to descriptors relies heavily on judgments that are often error-prone. For instance, student performance may be reported as the percentage of students who have attained "proficiency," though determining what constitutes proficiency may be a highly subjective process. Even so, assessment results continue to be reported in terms of cut scores, and it is therefore important to determine whether the cut scores set on different tests provide reasonably consistent

information about students. If a student is labeled "Advanced" or "Proficient" on a state test but is unable to reach the level of performance on a placement test necessary to avoid remedial coursework, there is reason to believe that the cut scores used on one or both tests are inappropriate.

The current report is limited in scope to the first aspect of alignment discussed above. Because we do not have access to test score data, we are not examining item statistics or relationships among scores on different tests and criterion measures (such as first-year grade point average). A comprehensive study of standards-setting across instruments is also beyond the scope of this project.

The Case for Appropriate Discrepancies among Tests

While alignment among different assessments is desirable in some instances, it is not always necessary or even warranted. Even when exams are designed to be parallel, as in alternate forms of the SAT, we would not expect perfect alignment. Because tests are merely samples from a larger universe of possible constructs, they are likely to differ from one testing occasion to the next. Thus, tests that are very similar in purpose, content, and format can rank order students in different ways. This underscores the importance of not placing too much emphasis on the rankings of one test, as no single exam can be a comprehensive or completely precise measure of a student's ability.

Moreover, there are instances in which discrepancies among the assessments are not only unavoidable, but also warranted. Below we describe three arguments in favor of maintaining discrepancies among assessments.

(1) Discrepancies among tests may result from appropriate efforts to tailor a measure to the situation for which it was designed.

When tests serve different purposes, they may necessitate different levels of cognitive sophistication, or vary in their sampling from the possible domain of constructs. A college entrance exam, for instance, should not be expected to resemble a remedial placement test. Because the college entrance exam is used to

select among higher-achieving students for entrance into universities and colleges, the tests need to include advanced content in order to distinguish among the examinees and rank order them consistently. On the other hand, a remedial placement test is used to determine whether examinees possess entry-level skills, and may therefore draw from more fundamental material. In this instance, the misalignments in content or skill levels stem from variations in the uses of the assessments, and eliminating the discrepancies between the two tests will undermine their intended purposes.

(2) Misaligned tests that provide multiple measures of an examinee's ability are needed for the validation process.

Even when exams serve similar purposes, discrepancies among such tests may actually be advantageous. First, it is virtually impossible for any test to measure all relevant aspects of a construct, and therefore, some knowledge and skills that are deemed important will necessarily be omitted. This problem is mitigated somewhat with the use of multiple tests that differ with respect to skills and content, as the discrepant exams can better capture the range of a given construct than well-aligned tests. Additionally, because tests differ on unintended as well as intended dimensions, having various achievement measures minimize the influence of idiosyncratic test features on performance. Two writing exams, for instance, that ask students for essays on the same content area makes it difficult to disentangle writing proficiency (e.g., intended construct) from familiarity with the particular topic (e.g., unintended construct). A more accurate measure of writing ability can be obtained if the writing prompts were to vary across the two assessments. In a similar vein, tests that share a common item format can potentially confound measurement method with performance. Although studies have shown that scores on achievement tests measuring the same general skills but differ in format tend to correlate highly, it is nevertheless desirable for students to be assessed in multiple ways, as strong correlations among the exams provide evidence for their validity (see Campbell and Fiske, 1959 for a discussion of convergent validity).

(3) Discrepant assessments provide a check against inappropriate test preparation.

Suppose scores on a K-12 exam were used to determine which students should graduate or which teachers should get bonuses in their paychecks. Due to the consequences attached to these scores, there is a significant risk of "teaching to the test." Empirical evidence has shown that under high-stakes circumstances, teachers may narrow their curriculum to the content that is found on the assessments, and give students undue instruction and practice on specific types of items that are likely to be tested (Stecher and Barron, 1999; Madaus, 1988). This type of preparation can lead to gains on the K-12 test, but the observed gains may not be associated with a commensurate increase in proficiency.

One way of detecting the problem of inappropriate test preparation is to compare scores on the K-12 test with scores on external assessments that measure the same general construct as the K-12 test, but vary with respect to content, skills, format, and context (e.g., low-stakes). Similar performances on the K-12 exam and these other discrepant measures provide evidence that the gains achieved on the K-12 test represent a substantive increase in proficiency. In contrast, large discrepancies in scores may indicate spurious gains, and raise important validity questions. This example illustrates the point made above: assessments varying in content and scope can facilitate score interpretations by providing multiple views of performance.

As noted above, completely aligned measures are neither likely nor desirable. Therefore, our discussions of misalignments should not be interpreted as a call for creating perfect consistency among different assessments. Instead, we focus on the signaling functions of the tests, noting the kinds of messages that students may receive.

National College Entrance Examinations across the Five States

The first set of tests we examined, which includes the SAT I, SAT II, ACT, and AP exams, are used nationally to aid in college admissions decisions. The SAT I, a

three-hour mostly multiple-choice exam is designed to measure general mathematical and verbal reasoning, is intended to help predict freshman grade point average in college. The SAT II is a one-hour multiple-choice test that assesses in-depth knowledge of a particular subject, and is used by admissions officers as an additional measure with which to evaluate student subject-matter competence. The SAT II is used primarily at the more selective institutions and is taken by far fewer students than is the SAT I. For this study, we examined the following SAT II tests: Mathematics IC, Mathematics IIC, Literature, and Writing. The ACT is an approximately three-hour exam consisting entirely of multiple-choice items. Used as an alternative measure to the SAT I in evaluating applicants' chances of success in college, it assesses achievement in several academic subjects, including science, reading, writing, and math. The AP tests are used to measure college-level achievement in several subjects, and to award academic credit to students who demonstrate college-level proficiency. We examined the Calculus AB and English Language AP exams.

Examinees are encouraged to take the ACT or SAT I within their junior or senior years, whereas the most optimal time to take the SAT II or AP exams is within months of completing a particular course. For those students applying to a four-year institution, many are required to take either the SAT I or ACT, and, at certain schools, several SAT II exams as part of the admissions process. While the AP tests are not a requirement, admissions officers are likely to view students with AP experience as better-prepared and more competitive applicants.

Other Kinds of Assessments

In addition to the college admissions tests, we also examine state-specific assessments, including the K-12 and placement exams. The K-12 exams are used to monitor student progress toward state standards, whereas the placement tests are used to place students in a course commensurate with their ability. Some institutions administer departmentally-created placement tests, whereas others rely on commercially available assessments.

The Oregon Assessment Environment

In the past ten years, Oregon has developed and implemented major policy shifts that affect both its K-12 and higher education systems. The most visible components of the K-12 reform efforts are the certificate of Initial Mastery (CIM) and Certificate of Advanced Master (CAM). These certificates are designed to be capstones to student mastery of the standards at the tenth and twelfth grades, respectively. Currently, only the CIM is available, as the CAM is still under development.

In order to earn a CIM, Oregon students are expected to meet standards based on performance on standardized assessments and in work samples in the tenth grade in English (reading and writing), math, speaking, science and social studies.[2] The CIM English test consists of a writing sample and a 65-item multiple-choice section. The CIM math test contains 55 multiple-choice questions, and one open-ended problem-solving item. The CIM English and Math assessments are level tests that provide more precise measures of proficiency by allowing students to take exams that are tailored to their achievement level. That is, depending upon ability, students take different versions of the CIM. To identify the appropriate test level for the students, teachers may use professional judgment and/or administer locator tests provided by the state. In English, the locator test is a 30-45 minute exam, consisting of 54 multiple-choice reading items. The math locator exam contains 24 multiple-choice questions administered within 40-minutes. We include all versions of the CIM as well as the locator tests for our study.

Placement Tests Used in Oregon

In addition to the CIM and CAM, students applying to a public university in Oregon may also be required to take placement tests in math and/or English. These tests are used to determine whether admitted students possess entry-level math and English skills. The colleges in Oregon administer a wide range of placement exams; we include the assessments used at the University of Oregon as an example. At the University of Oregon, students who do not meet the minimum achievement level on the SAT I or ACT are required to take the Test of Standard Written English (TSWE), which is a 30-minute, 50-item multiple-choice exam that assesses use of basic grammar, sentence structure, and

[2] In the spring of 2000, there was a change in CIM requirements. Rather than a single certificate for achievement across all subjects, individual certificates will be awarded by subject.

word choice. In math, all students must take a placement test except examinees with satisfactory scores on the AP Calculus exam, or those who have transferred credit for college-level calculus from another institution. The math placement test consists of 40 multiple-choice questions administered within 50 minutes. Scores on the placement test determines the mathematics course which students will be eligible to register for.

Tables 1 and 2 list these testing programs and the type of information we were able to obtain for this study. For most tests, we used a single form from a recent administration or a full-length, published sample test. In a few instances where full-length forms were unavailable, we used published sets of sample items. This was the case for the University of Oregon (UO) math placement test and Oregon state assessments. For the English/language arts (ELA) tests, the table specifies whether the test includes each of three possible types of items: reading, objective (e.g., multiple-choice) writing, and essay writing.

Table 1. Structural Characteristics of the Tests: Mathematics

Test	Materials Examined	Time Limit	Number of Items	Tools	Purpose	Framework	Content as Specified in Testing Materials
ACT	Full sample form	60 minutes	60 MC	Calculator	Selection of students for higher education	High school mathematics curriculum	Prealgebra (23%), elementary algebra (17%), intermediate algebra (15%), coordinate geometry (15%), plane geometry (23%) and trigonometry (7%)
AP Calculus AB	Full form, 1997 released exam	Two 90-minute sections	40 MC 6 Free response	Graphing calculator on last 15 MC items	Provide opportunities for HS students to receive college credit and advanced course placement	AP Calculus Course Description	Calculus
Certificate of Initial Mastery Mathematics Assessment (CIM)	Sample items	Two untimed testing sessions	55 MC 1 OE	Calculator	Monitor student achievement toward specified benchmarks	*Common Curriculum Goals* adopted by the State Board of Education	Calculations and estimations, measurement, statistics and probability, algebraic relationships, and geometry
Locator Test	Full sample form	40 minutes	24 MC	Calculator	Identify the appropriate form of CIM to be administered	*Common Curriculum Goals* adopted by the State Board of Education	Calculations and estimations, measurement, statistics and probability, algebraic relationships, and geometry
SAT I	Full sample form	Two 30-minute sessions	35 MC 15 QC 10 GR	Calculator	Selection of students for higher education	High school mathematics curriculum	Arithmetic (13%), algebra (35%), geometry, (26%), and other (26%)

		One 15-minute session					
SAT II-Level IC	Full sample form	60 minutes	50 MC	Calculator	Selection of students for higher education	Three-year college preparatory mathematics curriculum	Algebra (30%), geometry (38%, specifically plane Euclidean (20%), coordinate (12%), and three-dimensional (6%)), trigonometry (8%), functions (12%), statistics and probability (6%), and miscellaneous (6%)
SAT II-Level IIC	Full sample form	60 minutes	50 MC	Calculator	Selection of students for higher education	More than three years of college preparatory mathematics curriculum	Algebra (18%), geometry (20%, specifically coordinate (12%) and three-dimensional (8%)), trigonometry (20%), functions (24%), statistics and probability (6%), and miscellaneous (12%)
University of Oregon Math Placement Test	Full sample form	50 minutes	40 MC	Calculator	Placement of students into appropriate math course	University of Oregon Mathematics Departmental Standards and *Mathematics Association of America* Standards	Elementary algebra, intermediate algebra, geometry, precalculus

Notes.
MC = multiple-choice
OE = open-ended
GR = grid-in
QC = quantitative comparison

Table 2. Structural Characteristics of the Tests: English/Language Arts

Test	Materials Examined	Time Limit	Number of Items	Purpose	Framework	Reading Section?	Objective Writing Section?	Essay Section?
ACT	Full sample form	80 minutes (35 minute reading section, 45 minute objective writing section)	40 MC reading 75 MC objective writing	Selection of students for higher education	High school mathematics curriculum	Y	Y	N
AP Language and Composition	Sample questions	60 minute MC section, 120 minute essay section	52 MC 3 essays	Provide opportunities for HS students to receive college credit and advanced course placement	AP English Language and Composition Course Description	Y	N	Y
Certificate of Initial Mastery Reading Assessment	Sample form	No time limit	65 MC	Measure student achievement toward specified benchmarks	*Common Curriculum Goals* adopted by the State Board of Education	Y	N	N
Certificate of Initial Mastery Writing Assessment	Sample essays	No time limit	1 essay	Measure student achievement toward specified benchmarks	*Common Curriculum Goals* adopted by the State Board of Education	N	N	Y
Locator Test	Full sample form	30-45 minutes	54 MC	Identify the appropriate form of CIM to be administered	*Common Curriculum Goals* adopted by the State Board of Education	Y	N	N
SAT I	Full sample	Two 30-minute	78 MC	Selection of students	High school	Y	Y	N

	form	sessions One 15-minute session		for higher education	Reading and Language Arts curriculum			
SAT II-Literature	Full sample form	60 minutes	60 MC	Selection of students for higher education	High school English and American literature curriculum	Y	N	N
SAT II-Writing	Full sample form	One 40-minute MC session One 20-minute essay session	60 MC 1 essay	Selection of students for higher education	High school Reading and Language Arts curriculum	N	Y	Y
Test of Standard Written English (TSWE)	Full sample form	30 min	49 MC	Evaluate ability of college-bound students to recognize standard written English	High school Reading and Language Arts curriculum	N	Y	N

Alignment Among Oregon Math Assessments

In this section we describe the results of the alignment exercise for the math tests.[3] We first characterize the procedural and structural misalignments, then describe the cognitive and content discrepancies. Our math analysis concludes with a discussion of the implications, noting that some of the inconsistencies were unimportant, whereas other discrepancies were more serious, and still other misalignments were necessary.

Differences in the Procedural and Structural Aspects of the Alignments

The first aspect of the math framework concerned the technical dimension of the tests. This technical dimension involves those features of the test that could be described through simple examination of the test and items. This includes the number of items, time limit, format (e.g., multiple-choice, essay), provisions for the use of tools such as calculators or protractors, the use of diagrams or other graphics, the use of formulas, and whether each item was embedded in a context (as in a word problem). The use of formulas was sometimes difficult to determine because problems can be solved in multiple ways, and in some cases an item could be solved either with or without a formula. Items were coded as requiring a formula only if it was determined that the formula was necessary for solving the problem.

On this technical dimension, we found some similarities in the procedural and structural aspects of the assessments. The tests were alike in that they all included multiple-choice items, and most measures were administered in a single testing session. Students were usually allowed the use of a calculator, although most questions did not require extensive computation. They also assumed familiarity with basic formulas and mathematical identities as background for most of the questions, although knowledge of more complex formulas was seldom necessary.

However, the assessments differed in a number of ways. There was a great deal of structural variation among the exams, especially with regard to the percentages of items containing formulas and illustrations. Approximately 5% or less of the items on the SAT

[3] We did not include the results for the AP Calculus AB exam because it was markedly different from the other studied tests. For example, it did not include material from any other mathematical content area except calculus, and was the only measure that necessitated a graphing calculator. It was also intended to

I and UO math placement test required a memorized formula, in contrast to 16% of the questions on the CIM exam. Whereas the SAT II Level IIC and UO math placement exam made little use of figures (2% and 0%, respectively), the SAT II Level IC and CIM tests included many illustrations, with 26% of their questions containing a diagram. Differences in the degree to which tests require interpretation of spatial or figural information are particularly important as they can affect gender and other group differences.

assess the proficiency of a very select group of high-ability students, whereas the other assessments were intended for a wider range of ability levels.

Table 3: Structural, Content, and Cognitive Features of the Mathematics Tests

Test	Format				Context	Graphs			Diagrams			Formulas		Content								Cognitive Requirements		
	MC	QC	GR	OE	C	S	RO	P	S	RO	P	M	G	PA	EA	IA	CG	PG	TR	SP	MISC	CU	PK	PS
ACT	100	0	0	0	22	5	2	0	13	0	0	15	0	17	22	5	15	25	8	3	5	40	53	7
CIM	90	0	0	10	67	18	0	0	26	0	0	16	3	15	15	3	5	31	0	30	2	18	61	21
Locator test	100	0	0	0	60	20	0	0	20	0	0	15	0	25	5	0	15	20	0	30	5	20	75	5
SAT I	58	25	17	0	25	7	0	0	18	0	0	1	8	13	37	2	6	19	0	13	11	32	53	15
SAT II-Level IC	100	0	0	0	18	8	0	0	26	0	0	12	0	2	30	10	12	28	4	8	6	34	58	8
SAT II-Level IIC	100	0	0	0	12	12	2	0	2	0	0	10	0	2	14	22	12	14	18	6	12	26	54	20
University of Oregon Placement Test	100	0	0	0	8	0	0	0	0	0	0	5	0	0	65	15	5	3	13	0	0	10	90	0

Legend:

Format
MC = multiple-choice items
QC = quantitative comparison items
GR = fill-in-the-grid items
OE = open-ended items

Context
C = contextualized items

Graphs/Diagrams
S = graph/diagram within item-stem
RO = graph/diagram within response options
P = graph/diagram needs to be produced

Formulas
M = formula needs to be memorized
G = formula is provided

Content
PA = prealgebra
EA = elementary algebra
IA = intermediate algebra
CG = coordinate geometry
PG = plane geometry
TR = trigonometry
SP = statistics and probability
MISC = miscellaneous topics

Cognitive Requirements
CU = conceptual understanding
PK = procedural knowledge
PS = problem-solving

Differences in the Degree and Kinds of Cognitive Skills Required

This study also looked at the cognitive dimension of the tests. The cognitive dimension included conceptual understanding, procedural knowledge, and problem solving. As is typical with studies like this (e.g., Kenney & Silver, 1999), the raters found this dimension to be the most difficult to code, partly because items can often be solved in multiple ways, sometimes as a function of the examinee's proficiency. What might be a problem-solving item for one examinee might require another to apply extensive procedural knowledge. For instance, consider an item asking students for the sum of the first 101 numbers starting with zero. A procedural knowledge approach might involve a computation-intensive method, such as entering all the numbers into a calculator to obtain the resulting sum. However, the problem-solving approach would entail a recognition that all the numbers, except the number 50, can be paired with another number to form a sum of 100 (100+0, 99+1, 98+2, etc.). The total sum is then computed by multiplying the number of pairs (i.e., 50) by 100 and adding 50. Depending upon the chosen approach, the same item can elicit varying levels of mathematical sophistication.

Although an item cannot be unanimously classified as a conceptual, procedural, or problem-solving question, "what can be classified are the actions a student is likely to undertake in processing information and providing a satisfactory response." (NAGB, 1995). Using the same definitions as those used for NAEP, we coded the most probable course of action undertaken by a typical examinee who successfully answered the item. The descriptions of conceptual, procedural, and problem-solving are described below:

Table 4. Descriptions of the Cognitive Dimensions

Dimension	Definition
Conceptual understanding	Reflects a student's ability to reason in settings involving the careful application of concept definitions, relations, or representations of either
Procedural knowledge	Includes the various numerical algorithms in mathematics that have been created as tools to meet specific needs efficiently
Problem solving	Requires students to connect all of their mathematical knowledge of concepts, procedures, reasoning, and communication/representational skills in confronting new situations

Source: Mathematics Framework for the 1996 and 2000 National Assessment of Educational Progress (NAGB, 2000).

While it was difficult to neatly sort out the various cognitive functions, it was apparent that different reasoning requirements were elicited by different tests (see Table 3). Although none of the assessments focused heavily on problem-solving items, there were differences in the amounts and kinds of cognitive skills required. All of the items on the UO placement test and 95% of the locator test items entailed straightforward application of declarative and procedural knowledge. In a similar vein, the vast majority of questions on the ACT and SAT II Level IC tests necessitated that students apply an algorithm or heuristic to a textbook-like problem. On the other hand, the CIM and the SAT II Level IIC, the latter intended for examinees enrolled in more advanced college preparatory math courses, placed the most emphasis on problem-solving ability. Approximately 20% and 21% of the items on the SAT II Level IIC and CIM, respectively, assessed higher-order thinking skills.

The SAT I and CIM were similar in that both included items that required students to generate their own responses. The open-ended questions on the CIM, however, were much more extensive than the SAT I items. Whereas the SAT I open-ended items generally did not call for multiple strategies, successful solution of a CIM open-ended problem generally required multiple steps, and students were asked to justify or explain their solutions. Thus, although both tests made use of a free-response format, the cognitive demands differed dramatically.

Differences in Content Tested

The content dimension of the tests was also examined. This content dimension included several categories of math topics, from pre-algebra (e.g., numbers and operations) through calculus. Almost all of the tests we examined had specifications that included many or all of these categories. We listed sub-categories as means of making the distinctions among the main categories clearer, but we coded only using the main categories.

Although all the exams are considered measures of mathematics achievement, there was some variation in the constructs assessed. Approximately 65% of the UO math placement test measured elementary knowledge, whereas the CIM focused more on plane geometry and statistics (31% and 30% of its items, respectively). For college

admissions exams such as the SAT II Level IIC and ACT, relatively greater emphasis was given to trigonometry, a topic that was absent from the CIM, locator, and SAT I.

Variations in Contextualization

Instances of discrepancies were also observed with respect to the amount of contextualization provided. Although many curricular experts in math have suggested that test items be set within the context of everyday life, most of the exams continued to measure student achievement with abstract questions. No more than 25% of the items found on the college admissions and placement assessments were contextualized, and as few as 12% of the items (e.g., SAT II Level IIC) were embedded in a real-life situation. Furthermore, the UM placement test was completely devoid of any contextualized items. These assessments are in stark contrast to the Oregon state assessments, where 67% of the CIM items and 60% of the locator test items were framed in realistic situations.

Discrepancies among the Frameworks

There were also variations in the frameworks that guided the development of the tests. The CIM and locator exams followed state-mandated standards (specifically, *Common Curriculum Goals*) that contained detailed learning objectives. The guidelines for the other assessments, such as the UO placement test, ACT and SAT IIs were more loosely defined, partly because they were not based solely on Oregon's high school math curricula. These frameworks also placed relatively greater emphasis on content found in more advanced math courses. The SAT I guidelines, on the other hand, were less curriculum-based than the frameworks for the other assessments, and instead focused on general mathematical reasoning abilities developed over years of schooling.

Implications of the Discrepancies among the Math Assessments

In many instances, the math assessments examined were discrepant with respect to format, content, and cognitive requirements. Some of the inconsistencies were trivial, whereas others were potentially more serious. Still other discrepancies were not only inevitable, but also desirable. Below, we discuss the implications of the misalignments noted earlier.

Some Discrepancies are Unimportant

Some of the misalignments may not merit much concern, as they are unlikely to affect student perceptions of the kinds of skills that are considered important, or have any other discernible negative effect on students' preparation efforts. Variations in time limits, for instance, would most likely fall within this category.

Some Discrepancies Represent a Gap between what is Seen as Important and what is Actually Included on the Tests

Despite efforts to align the assessment to its framework, there were instances in which the curricular task did not match its guidelines. For example, although their frameworks emphasized contextualization, few items on the ACT, SAT I, and SAT II were framed in a real-life setting. This dearth of contextualized questions may lead students to perceive math as bearing little relevance to their everyday lives outside of the school or testing context.

In a similar manner, although the *Mathematics Association of America* standards emphasized mathematical problem-solving and communication, the exam resulting from this frameworks was not well aligned with these stated goals. None of the items on the UO placement test assessed problem-solving ability. The multiple-choice format favored by this assessment can also send potentially negative messages regarding the importance of reasoning skills. Although items in any format can be designed to measure a variety of abilities, multiple-choice items are popularly believed to be less adequate than free-response questions at measuring higher-order thinking (Frederiksen, 1984). Furthermore, multiple-choice items do not require students to communicate mathematically, as students who select the correct option receive full credit, and do not have to justify or explain their logic or thought processes underlying the given response.

Sometimes Discrepancies are Warranted and Necessary

When discrepancies emerge from appropriate efforts to adapt a test to serve a particular purpose, the differences may be unavoidable. For instance, although the SAT Level IIC included topics from a wide variety of courses, including trigonometry, the

locator exam did not include material beyond geometry. In this particular case, the SAT Level IIC and locator exam have disparate purposes, and content differences are therefore to be expected. Because the SAT Level IIC is typically used by more selective universities as part of the admissions process, it needs to include topics from advanced courses to determine whether incoming students have the requisite background. The locator test, on the other hand, is intended to be a measure of fundamental mathematics knowledge. Consequently, it is more appropriate for this assessment to limit its content to a narrow area of math than to sample extensively from the entire mathematics domain.

Even when assessments serve similar purposes, some misalignments can be valuable. Consider the ACT and SAT I, which are often used interchangeably for college admissions, yet contain different item types, and require students to demonstrate substantially different skills and knowledge. Given the high-stakes nature of these assessments, it is desirable to have the measures differ somewhat, so as to check against artificial inflation of scores resulting from inappropriate coaching. Additionally, because discrepant measures provide broader coverage of the intended domain than well-aligned assessments, the inconsistencies between the ACT and SAT I can actually be advantageous, as they provide a more comprehensive indicator of examinees' abilities.

Finally, some of the discrepancies noted above are indicative of variations in curricular frameworks. The *Common Curriculum Goals*, for example, placed more emphasis on contextualization than did the frameworks for the other tests; as a result, the CIM and locator exams contained the greatest proportion of questions framed within realistic settings. Likewise, content discrepancies between the ACT and SAT I were reflective of their framework variations. Because the specifications for the ACT were more closely tied to the core high school curriculum than were the guidelines for the SAT I, the ACT contained more textbook-like items, whereas the SAT I contained more novel questions.

Misalignments in Oregon English Language Arts (ELA) Assessments

In this section we present the results of our analysis of alignment among tests used to assess students' skills in reading and writing. The tests' names were varied, but they all focused on reading and/or writing in the English language. Table 2, discussed

briefly above, lists the tests along with basic details. Below we discuss the ELA framework.

Framework

The ELA framework covered three types of items: reading, objective writing (mainly multiple-choice items), and essay writing. Many of the tests we examined included two or all three of these item types, whereas others focused on a single type. In contrast to mathematics, there were no clear content areas that could be used to categorize items. Instead, the ELA analysis focused more on structural characteristics and cognitive demands. In addition, many of the tests included short passages followed by sets of items, so it was necessary to categorize both the passage and the individual item.

There was extensive overlap among the frameworks for reading, objective writing, and essay writing. As with math, we identified subcategories to sharpen the distinctions among the main categories, but we coded using only the main categories. The structural dimensions, described in further detail in Table 5a, included three categories. The topic category captured the subject matter of the passage, and consisted of five areas—fiction, humanities, natural science, social science, or personal accounts. The type category identified the author's writing style as narrative, descriptive, persuasive, or informative. The stimulus category referred to the presentation of the passage, such as a letter, essay, poem, or story. Raters used all three categories when coding the reading and objective writing items, but used only the topic category when coding the essay writing questions.

The cognitive framework for both the reading and objective writing measures consisted of three dimensions. Raters coded questions as assessing ability to (i) recall information, (ii) make inferences, or (iii) evaluate an item's style. In reading, questions that could be answered via direct reference to the passage were coded as recall items, whereas questions that required the examinees to interpret the material were coded as inference items. Questions that pertained to the development of ideas or improved upon the presentation of the reading passages were coded as evaluating style.

For the objective writing measures, items that entailed application of grammatical rules were considered recall items. Typically, most of these questions concerned mechanics or usage errors. Inference items were those that required examinees to identify cause-and-effect relationships, and evaluating style items related to rhetoric ability, such as sentence organization, clarity, and other stylistic skills. Table 5b gives more details of the cognitive coding systems.

The above framework was not applicable to the essay writing items, so different guidelines were developed (see Table 5b). For the essay writing questions, raters focused on the scoring criteria, particularly the emphasis given to mechanics, word choice, organization, style, and insight.

Table 5a. Description of the ELA Structural Dimension Coding Scheme

Dimension	Description or Example
Type of Writing	
Narrative	Stories, personal essays, personal anecdotes
Descriptive	Describes person, place, or thing
Persuasive	Attempt to influence others to take some action or to influence someone's attitudes or ideas
Informative	Share knowledge; convey messages, provide information on a topic, instructions for performing a task
Topic	
Fiction	story, poem
Humanities	e.g., artwork of Vincent Van Gogh
Natural sciences	e.g., the reproductive process of fish
Social sciences	e.g., one man, one vote; cost effectiveness of heart transplants
Personal	e.g., diary account of death of a parent
Stimulus materials	
Letters	
Essays	
Poems	
Stories	

Table 5b. Description of the ELA Cognitive Dimension Coding Scheme

Dimension	Description or Example	Used for Reading	Used for Objective Writing	Used for Essay Writing
Cognitive Demands		X	X	
Recall	Answer can be found directly in the text, or by using the definitions of words or literary devices, or by applying grammatical rules	X	X	
Infer	Interpret what is already written	X	X	
Evaluate style	Improve the way the material is written	X	X	
Scoring Criteria				X
Mechanics	Grammar, punctuation, capitalization			X
Word choice	Use of language, vocabulary, sentence structure			X
Organization	Logical presentation, development of ideas, use of appropriate supporting examples			X
Style	Voice, attention to audience			X
Insight	Analytic proficiency, accurate understanding of stimulus passage, thoughtful perceptions about its ramifications			X

Differences in Test Structure and Administration

As with math, differences among the exams were prevalent. Some assessments did not involve a written composition (ACT, SAT I, SAT II Literature and locator test), whereas others required multiple essays (AP). There were also vast differences in the amount of time students were permitted to write their compositions; the AP allowed two hours to compose three essays, whereas the SAT II Writing exam allowed only 20 minutes.

The discrepancies were not limited to the administrative characteristics of each exam, but were also apparent with respect to the structural features. In reading, all of the passages on the SAT II Literature test were narrative, and 63% were on fictional topics (see Table 6a). In contrast, the SAT I passages tended to be informative (60%), and were much more likely to draw from humanities (40%). The essay was the most predominant presentation mode, with all of the passages on the AP exam presented in this manner. The

majority of the passages on the SAT I, ACT, and locator test were also essays (80%, 75%, and 75% respectively), but the SAT II Literature and the CIM varied the stimuli in which the reading passages were presented. The SAT II Literature exam tended to use poems as item prompts (50%), whereas the CIM used stories (50%).

On measures of objective writing, the ACT and SAT II Writing assessments included passages as an item prompt, whereas the SAT I and TSWE did not (instead, they used short sentences). There was little evidence of misalignment with respect to stimulus or type, as every passage was presented as an essay, and was either narrative or informative (see Table 6b). There was also little variation in topics; all the SAT II Writing items drew upon humanities, as did the majority of the ACT questions.

Table 6a: Percent of Reading Passages Falling into Each Category

Test	Type				Topic					Stimulus				
	Narrative	Descriptive	Persuasive	Informative	Fiction	Humanities	Natural Science	Social Science	Personal	Letter	Essay	Poem	Story	Chart
ACT	50	0	0	50	25	25	25	25	0	0	75	0	25	0
AP	75	0	0	25	0	25	25	0	50	0	100	0	0	0
CIM Reading	75	0	0	25	63	13	0	0	25	0	13	13	50	25
Locator Test	75	0	0	25	25	0	25	25	25	0	75	25	0	0
SAT I	40	0	0	60	20	40	20	20	0	0	80	0	20	0
SAT II Literature	100	0	0	0	63	0	0	13	25	13	25	50	13	0

Table 6b: Percent of Objective Writing Passages Falling into Each Category

Test	Type				Topic					Stimulus			
	Narrative	Descriptive	Persuasive	Informative	Fiction	Humanities	Natural Science	Social Science	Personal	Letter	Essay	Poem	Story
ACT	40	0	0	60	0	60	20	0	20	0	100	0	0
SAT I	0	0	0	0	0	0	0	0	0	0	0	0	0
SAT II Writing	50	0	0	50	0	100	0	0	0	0	100	0	0
TSWE	0	0	0	0	0	0	0	0	0	0	0	0	0

For the extended essay writing assignments, the topic contents did not vary greatly from one test to the next (see Table 6c). The AP drew upon personal essays and humanities, whereas the CIM included personal essays and fiction. The SAT II Writing included themes from humanities.

Table 6c. Topic Contents of Essay Writing Prompts

Test	Topic				
	Fiction	Humanities	Natural Science	Social Science	Personal Essay
AP		X			X
SAT II Writing		X			
CIM Writing	X				X

Discrepancies in how Reading and Verbal Skills are Assessed

There were some discrepancies in terms of the cognitive demands the reading measures placed upon students (see Table 7a). Of the reading assessments we examined, only the AP exam required students to analyze the literary excerpts via a written composition. The remaining tests assessed knowledge and understanding of a passage solely with multiple-choice items. The SAT I and SAT II Literature tests placed great emphasis on analytical ability, with 83% and 80% of their items, respectively, assessing inferential skills (see Table 7a). To a lesser extent, the locator test also emphasized inferential skills (55%). In contrast, the ACT and CIM focused on straightforward recollection of information (58%, and 54% of their questions, respectively).

Table 7a. Percent of Reading Items Falling into Each Category

Test	Recall	Infer	Evaluate Style
ACT	58	42	3
AP	23	77	0
CIM Reading	54	46	0
Locator Test	45	55	0
SAT I	18	83	0
SAT II Literature	13	80	7

As was the case with math, two reading tests may have the same construct label, yet make vastly different cognitive demands. The CIM, AP Literature and Composition, SAT II Literature test, and ACT are all measures of reading proficiency, but differ in the kinds of skills assessed. The ACT items typically entailed recollection of facts directly from a given passage, and usually did not ask students to judge the mood or tone of the piece. The AP, SAT II Literature, and CIM assessments, on the other hand, required deeper analysis of the reading passage, oftentimes asking students to determine the effect of a given line or infer the intentions of the author. The AP and CIM, in particular, required students to apply their knowledge of literary devices. The AP and CIM tests included many items asking students to identify examples of hyperboles, alliterations, allusions, and the like, but such questions were not found on either the ACT or the SAT II Literature exams.

Discrepancies in How Writing Skills are Assessed

Inconsistencies with respect to the cognitive demands were also evident among the objective writing assessments (see Table 7b). Of the measures, only the SAT I included a significant proportion of items assessing inferential skills (100%). Such questions comprised less than 5% of the items on the ACT and SAT II Writing exams, and were completely absent from the TSWE. Instead, the TSWE focused on recall items. (90%). On the other hand, the ACT and SAT II Writing tests were more balanced in the kinds of skills they assessed; the items on these tests were mainly divided among recollection of information and evaluation of style.

Table 7b: Percent of Objective Writing Items Falling into Each Category

Test	Recall	Infer	Evaluate Style
ACT	48	4	48
SAT I	0	100	0
SAT II Writing	50	3	47
TSWE	90	0	10

There was much more consistency with respect to the kinds of cognitive demands required by measures of writing ability (see Table 7c). Skills such as mechanics, word

choice, style, organization, and insight were identified as important factors in virtually all of the tests we studied. However, the SAT II Writing test did not identify insight, which is the accurate understanding of the stimulus passage and thoughtful perceptions about its ramifications, as part of its scoring criteria, and the CIM gave more emphasis to grammatical rules than did the other assessments (see Table 6).

Table 7c. Factors Identified in the Scoring Criteria of Each Test

Test	Scoring Criteria Factors				
	Mechanics	Word Choice	Organization	Style	Insight
AP	X	X	X	X	X
CIM Writing	X	X	X	X	X
SAT II Writing	X	X	X	X	

Implications of the English Language Arts Discrepancies

As was the case with math, we observed discrepancies that were trivial, potentially negative, and unavoidable and necessary. Below we discuss the implications of the ELA inconsistencies in more detail.

Some Misalignments are Unimportant

Discrepancies that do not appear to influence student behavior or reflect legitimate differences stemming from different purposes are of little concern. These kinds of inconsistencies include minor differences in time limits, and variations with respect to whether or not the objective writing measures contained a reading passage.

Some Discrepancies Send Mixed Signals Regarding What Students Should Learn

Although most ELA assessments used similar scoring criteria, the omission of insight from the SAT II Writing scoring guidelines raised some concerns. First, this skill is part of the scoring criteria in most English courses, and for the other assessments we examined. This means that the SAT II Writing standards are incongruent with those that are typically expressed. Additionally, when scoring the test, raters are likely to be concerned with this factor, as the correct interpretation of the item prompt is oftentimes

considered an earmark of a well-written composition. If the SAT II Writing raters are tacitly including this skill as part of the scoring criteria, then students have not been provided with clear guidelines on how their writing is judged. In light of the kinds of signals the scoring rubric send, developers of the SAT II Writing assessments may wish to reconsider the current scoring criteria.

Some Discrepancies Represent a Mismatch between what is Valued and what is Tested

There may also be some inconsistencies between the skills that are valued and the skills that students are asked to demonstrate. Despite the importance of writing skills in many university courses, the two assessments that are most commonly used for the college admissions process, the SAT I and ACT, do not require students to write essays, and instead, assess writing ability with the multiple-choice format. This has important implications for the kinds of writing skills that are emphasized; aspects of writing proficiency that are less amenable to measurement with the multiple-choice format (e.g., personal writing style) tend to be de-emphasized or ignored. Most colleges and universities, however, value these kinds of skills, and tests that are used for the admissions process should ideally reflect this message.

Discrepancies between the curricular standards and the actual achievement tasks were also apparent. For instance, the ability to learn the meaning of a word from context is perceived to be an integral aspect of English, yet most of the tests did not truly assess this skill. Instead, many of the vocabulary items assessed students' recall ability rather than their inferential skills. The ACT, AP, CIM, and SAT II Literature assessments typically framed a vocabulary item as follows: "In lines XX, the word "panacea" is best understood to mean..." Although the question is phrased to indicate that the meaning relies on context, it can be construed as a recall question, as *a priori* knowledge of the definition is sufficient for a correct answer, since the context of lines XX did not affect the standard definition of "panacea."[4]

[4] As discussed earlier, whether an item assesses inferential skills or recall ability depends upon a students' proficiency level.

Some Misalignments Should be Maintained

As in math, some misalignments among the ELA assessments were necessary. Consider, for instance, the discrepancies between the scoring standards of the CIM and AP exam. For the former test, maximum scores were awarded to sample essays that had mechanics lapses, underdeveloped paragraphs, and diction errors. Under the AP guidelines, such compositions might receive adequate scores, but would not be viewed as an exemplary paper; only essays that demonstrate exceptional rhetorical and stylistic techniques would receive a maximum score under the AP scoring rubrics. Because the AP exam is intended for a select group of highly prepared students, whereas the CIM is taken by examinees of more moderate abilities, discrepancies between their scoring criteria are inevitable, as they reflect differences in the student population served by the tests.

In other instances, the discrepancies are probably avoidable, but should be maintained for validity purposes. This is the case for inconsistencies with respect to passage type, topic, and stimulus. Tests with little or no variation in these areas would not only encourage a narrowing of the curriculum, but interpretations regarding examinees' abilities would be confounded by construct-irrelevant aspects, such as knowledge of or familiarity with the passage topic. Studies have shown that different passages or topics can elicit vastly different performances, partly because interest differences can inappropriately affect scores (Carlton & Harris, 1992). It is for this reason that item prompts within and across different tests typically contain multiple passages that span a wide array of interest areas, and employ various kinds of stimuli and writing techniques.

From a validity perspective, discrepancies in emphases of particular cognitive skills are also warranted. Consider the objective writing section of the SAT I. Its heavy emphasis on inference skills means that other important abilities, including evaluative skills and recall ability, are not well measured. If the other objective writing measures were very well-aligned with the SAT I, the latter skills would have been ignored, and validity questions might arise, as the set of tests would represent a very narrow range of the intended domain. However, assessments and measures (e.g., transcripts) that are discrepant to the SAT I can capture other aspects of the intended construct. This example

underlies the importance of why high-stakes decisions should not be made on the basis of a single measure, as any particular test can assess only limited aspects of the intended domain.

Conclusion: Some Discrepancies are Problematic, Whereas Other Misalignments are Inevitable and Warranted

In general, many of the studied tests were not well-aligned with respect to structure or content. Some of these discrepancies were problematic, as they sent confusing signals to students regarding the kinds of skills that are valued. Other inconsistencies reflected legitimate differences stemming from diverse uses of the measures, or were necessary for validity purposes. Below we summarize some of the implications:

(1) Discrepancies can lead to a dissociation between the skills that are considered valuable and the skills that are actually assessed.

Although most mathematical testing frameworks emphasized problem-solving and communication as part of their measurement goals, the majority of the items on the studied assessments relied on routine items that did not require students to explain or justify their responses. Furthermore, despite reform efforts calling for an increase in the number of items framed in a more realistic setting, many items continued to be presented in an abstract manner. This lack of contextualization is problematic, as it may suggest to students that mathematics has little relevance to real-life problems.

Similarly, on the ELA assessments, students are not given clear signals as to which skills are valued. Many of the exams administered at the high school level, including the two tests most commonly used for college admissions (ACT and SAT I), do not require examinees to demonstrate their writing skills. The SAT II Writing test (which does include an essay item) does not require multiple writing samples, nor does it allow an extended period of time for students to fully develop their ideas in a single essay. Given that most college-level courses require students to write extensively, there is a discrepancy between the kinds of skills required in universities and those

required by the college admissions tests. Again, if the measures are to send signals that writing ability is a desired skill, then the current tests need to be modified to reflect that message.

(2) Discrepancies that stem from different purposes are not necessarily problematic. Oftentimes, tests with different uses will necessitate variations in content and skill levels that are contradictory, but warranted. The narrow content sampling of a course-specific exam should not be broadened to match that of a statewide achievement test. Similarly, trigonometry items that are inappropriate for a remedial placement test may be justifiable on an AP exam. These kinds of discrepancies should be maintained, as they reflect efforts to tailor the test for a particular purpose.

(3) Some misalignments are needed for validity purposes. Because it is virtually impossible for any given test to represent all of the pertinent knowledge and skills, well-aligned assessments may actually be a disadvantage, as they are likely to capture a smaller range of the intended domain than discrepant measures. Well-aligned tests also exacerbate the problems of irrelevant difficulty, as the same (unintended) factors that affect scores on one test are also likely to influence performance on the other, well-aligned measures. For example, if all the reading measures were to assess reading comprehension with passages pertaining only to solar systems, it would be difficult to interpret scores, as both reading ability (intended construct) and familiarity with solar systems (unintended construct) are likely to influence performance. By varying the content across assessments, the influence of non-construct related factors can be minimized, and scores can better reflect proficiency.

Finally, the threat of inappropriate test preparation on high-stakes tests calls for other measures to which scores can be compared. High-stakes exams often encourage a narrowing of the curriculum, and having external assessments that serve as another indicator of performance facilitates interpretations regarding the generalizability of any score gains observed on the high-stakes measures.

Although numerous discrepancies among the assessments were observed, only a few were considered problematic. Recommendations regarding what should be done about the inappropriate misalignments are beyond the scope of this report. However, it seems apparent that policymakers should attempt to address the potentially negative signals that these discrepancies send, possibly by better aligning the assessments to their frameworks or to each other. The latter course of action, however, should be judiciously undertaken, as inappropriate alignment can undermine the validity and usefulness of a given test. Alternatively, changing the tests may not be necessary, as long as examinees are provided with clear guidelines that enable them to prepare appropriately for each test and to know what will be expected of them at each step in the college admissions and placement process.

References

Achieve, Inc. (1999). *National Education Summit Briefing Book* (available at http://www.achieve.org).

American Federation of Teachers (1999). *Making standards matter 1999.* Washington, DC: Author.

Burstein, L., Koretz, D., Linn, R., Sugrue, B., Novak, J., Baker, E., & Harris, E. L. (1995/1996). Describing performance standards: Validity of the 1992 National Assessment of Educational Progress achievement level descriptors as characterizations of mathematics performance. *Educational Assessment, 3,* 9-51.

Carlton, S.T., & Harris, A.M. (1992). *Characteristics associated with differential item performance on the Scholastic Aptitude Test: Gender and majority/minority group comparisons* (ETS-RR-92-64). Princeton, NJ: Educational Testing Service.

Frederiksen, N. (1984). The real test bias. *American Psychologist, 39,* 193-202.

Madaus, G.F. (1988). The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical issues in curriculum* (pp. 83-121). Chicago: University of Chicago Press.

National Academy of Education (1993). *Setting performance standards for student achievement: A report of the National Academy of Education Panel on the evaluation of the trial state assessment: An evaluation of the 1992 achievement levels.* Stanford, CA: The National Academy of Education.

National Assessment Governing Board (1995). *Mathematics Framework for the 1996 National Assessment of Educational Progress: NAEP Mathematics Consensus Project.* (GPO 065-000-00772-0). U.S. Department of Education (available at http://www.nagb.org).

National Assessment Governing Board (2000). *Mathematics Framework for the 1996 and 2000 National Assessment of Educational Progress: NAEP Mathematics Consensus Project.* U.S. Department of Education (available at http://www.nagb.org).

National Assessment Governing Board (1995). *Mathematics Framework for the 1996 and 2000 National Assessment of Educational Progress.* U.S. Department of Education (available at http://www.nagb.org).

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5,* 299-321.

Smith, M., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233-268). Bristol, PA: The Falmer Press.

Stecher, B. M., & Barron, S. I. (1999). Quadrennial milepost accountability testing in Kentucky (CSE Report 505). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.